dbGaP: Database of Genotype and Phenotype

A collection of data from genome wide association studies and other clinical studies

https://www.ncbi.nlm.nih.gov/gap/

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

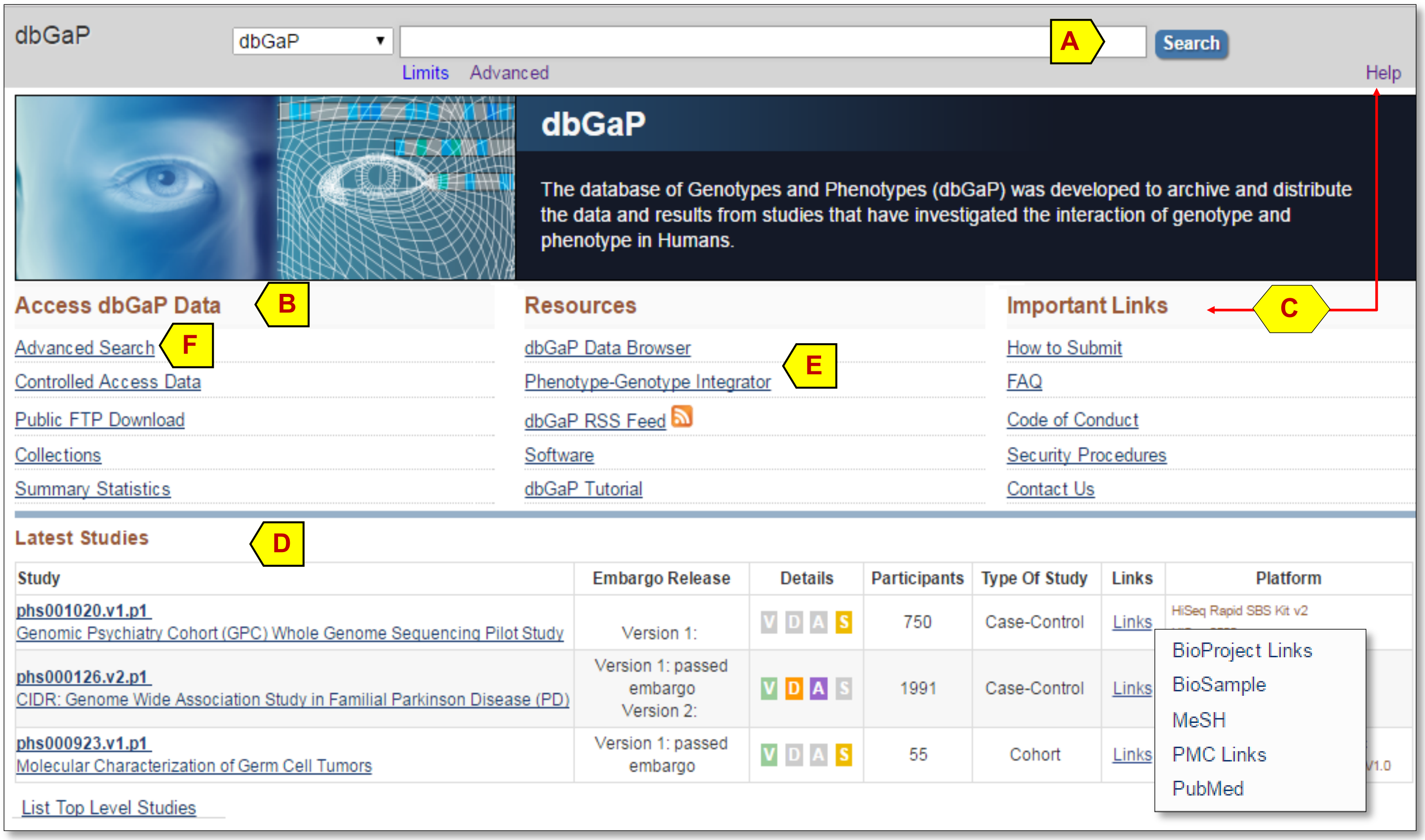
Scope and Access

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. Such studies include genome-wide association studies (GWAS), medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits. The advent of high-throughput, cost-effective methods for genotyping and sequencing has provided powerful tools that allow for the generation of the massive amount of genotypic data required to make these analyses possible.

dbGaP provides two levels of access. Open access through the dbGaP homepage (https://www.ncbi.nlm.nih.gov/gap) provides the public with summaries of studies, the contents of measured variables as well as original study document text. Access to individual-level data, including phenotypic data tables and genotypes, requires varying levels of authorization. Information on controlled access is available at: https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login. Access requires an eRA Commons login.

dbGaP Homepage

The homepage of dbGaP (shown below) provides a central entry point to access the data from this database. Entering terms in the search box and clicking the Search button (**A**) performs a search against this database. The Access dbGaP Data section (**B**) lists links to browse and access public content from the database or apply for controlled access to get the detailed data. Important Links and help (**C**) provides dbGaP-specific documentation and help. The Latest Studies (**D**) presents a selective list of GWAS dataset recently deposited to dbGaP with their titles linking to corresponding entries, and additional summary information given in the columns to the right. Integrated searches for phenotype and genotype data can be done using the Phenotype-Genotype Integrator (**E**).



The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

Study	Embargo Release	Details	Participants	Type Of Study	Links	Platform
phs001020.v1.p1 Genomic Psychiatry Cohort (GPC) Whole Genome Sequencing Pilot Study	Version 1:	V D A S	750	Case-Control	Links	HiSeq Rapid SBS Kit v2
phs000126.v2.p1 CIDR: Genome Wide Association Study in Familial Parkinson Disease (PD)	Version 1: passed embargo Version 2:	V D A S	1991	Case-Control	Links	
phs000923.v1.p1 Molecular Characterization of Germ Cell Tumors	Version 1: passed embargo	V D A S	55	Cohort	Links	V1.0

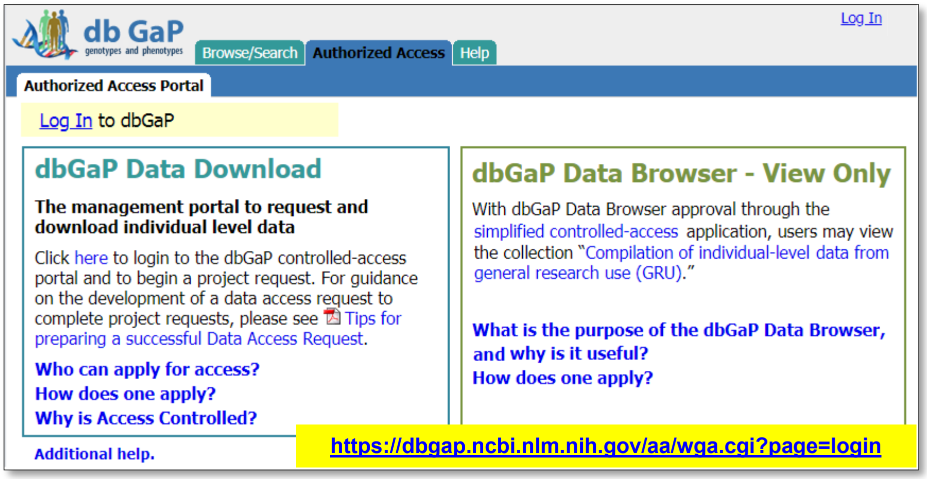
Legend for colored icons under the Details column: ***V****=variable,* ***D****=documents,* ***A****=analysis, and* ***S****=SRA data.*

The Advanced Search link (**F**) provides an alternative way to find studies of interest. It offers many sets of filters to facet existing studies into various categories to allow quick identification of studies of interest. For more information, see this webinar from the NCBI YouTube channel: https://www.youtube.com/watch?v=ePQ9p2SL_wM

Data Available from dbGaP and Controlled Access

Data deposited in dbGap come from various types of GWAS. These include longitudinal, case control, and cohort studies. For phenotype, the data include information collection forms, description of phenotypes, standards of measurement, and details of the individual phenotypes. For genotype, the data includes genotype calls and their quality scores from various platforms. If sequencing and expression analyses are included, the data will be brokered through other NCBI databases such as Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA). The summary of the phenotype and genotype data, associated through appropriate statistical methods, plus the analysis details are also available. Two levels of data access, public and controlled, are adopted to protect the privacy of study participants. Data that can potentially be used to establish personal identity of the participants are placed under restricted access. These include individual phenotypes, genotypes, sequence reads, expression profiles, epigenetic markers and full result sets. Information on the data collection, standards of phenotype observation and measurements, platforms for genotype determination, and the final summary of association analyses are available to the public through the dbGaP homepage.

For validated research needs with institutional support, a principal investigator can apply for controlled access to a specific dataset using the page shown to the right. Once a request is granted, the necessary information and credentials will be provided to the applicant so the specific dataset can be downloaded for further analysis. Further assistance is available by contacting dbGaP help: dbgap-help@ncbi.nlm.nih.gov



Log In to dbGaP

dbGaP Data Download

The management portal to request and download individual level data

Click here to login to the dbGaP controlled-access portal and to begin a project request. For guidance on the development of a data access request to complete project requests, please see Tips for preparing a successful Data Access Request.

Who can apply for access?
How does one apply?
Why is Access Controlled?

Additional help.

dbGaP Data Browser - View Only

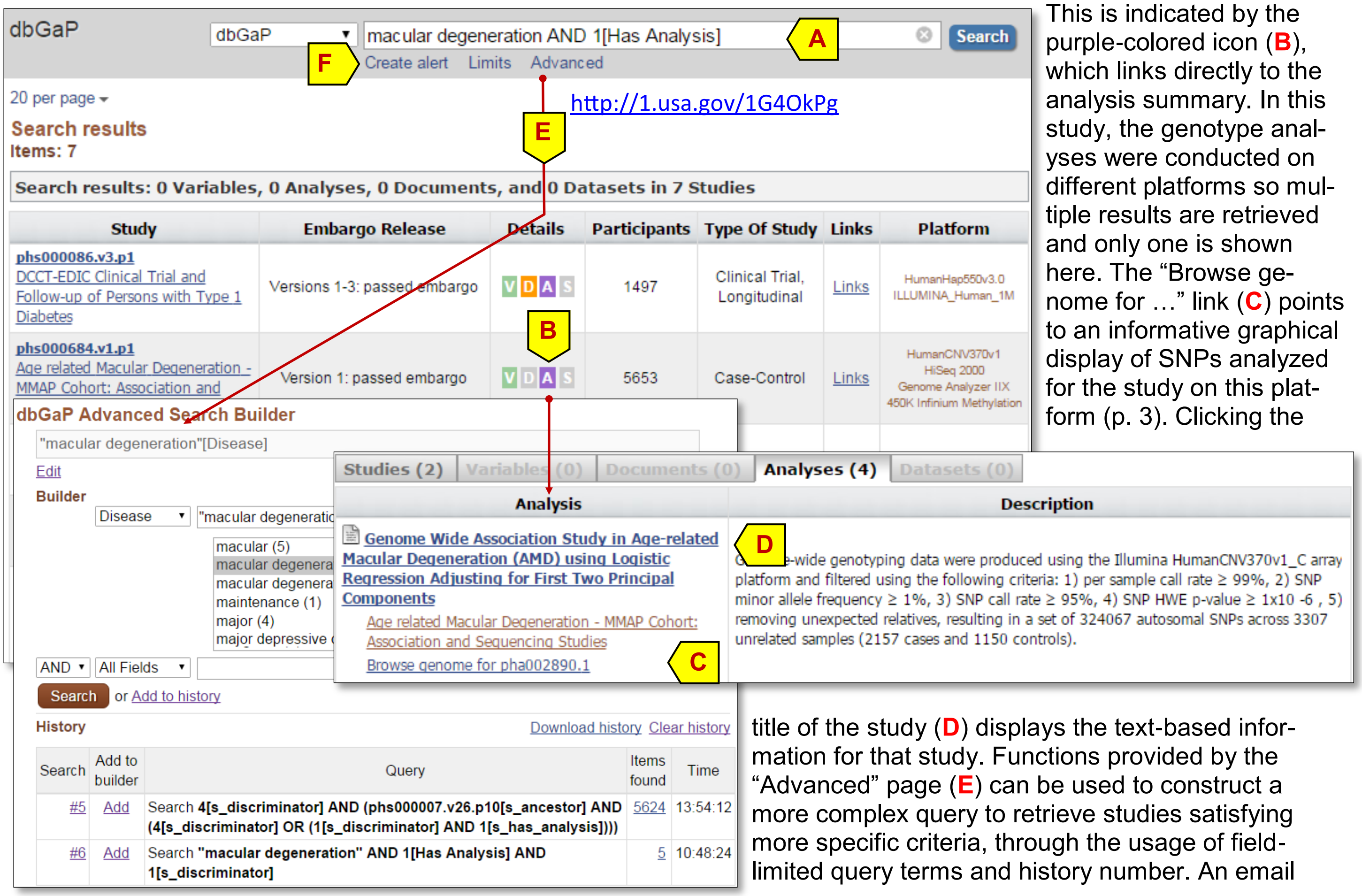
With dbGaP Data Browser approval through the simplified controlled-access application, users may view the collection "Compilation of individual-level data from general research use (GRU)."

What is the purpose of the dbGaP Data Browser, and why is it useful?
How does one apply?

https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login

Searching in dbGaP

Data available for public access can be searched and retrieved through the dbGaP homepage. In the example below, the disease "macular degeneration" is combined with filter "1[has analysis]" to retrieve related studies with full analysis (A). This is indicated by the purple-colored icon (B), which links directly to the analysis summary. In this study, the genotype analyses were conducted on different platforms so multiple results are retrieved and only one is shown here. The "Browse genome for …" link (C) points to an informative graphical display of SNPs analyzed for the study on this platform (p. 3). Clicking the title of the study (D) displays the text-based information for that study. Functions provided by the "Advanced" page (E) can be used to construct a more complex query to retrieve studies satisfying more specific criteria, through the usage of field-limited query terms and history number. An email



macular degeneration AND 1[Has Analysis]

http://1.usa.gov/1G4OkPg

Search results
Items: 7

Search results: 0 Variables, 0 Analyses, 0 Documents, and 0 Datasets in 7 Studies

Study	Embargo Release	Details	Participants	Type Of Study	Links	Platform
phs000086.v3.p1 DCCT-EDIC Clinical Trial and Follow-up of Persons with Type 1 Diabetes	Versions 1-3: passed embargo	V D A S	1497	Clinical Trial, Longitudinal	Links	HumanHap550v3.0 ILLUMINA_Human_1M
phs000684.v1.p1 Age related Macular Degeneration - MMAP Cohort: Association and	Version 1: passed embargo	V D A S	5653	Case-Control	Links	HumanCNV370v1 HiSeq 2000 Genome Analyzer IIX 450K Infinium Methylation

dbGaP Advanced Search Builder

"macular degeneration"[Disease]

Edit

Builder

Disease ▾ "macular degeneratic

macular (5)
macular degenera
macular degenera
maintenance (1)
major (4)
major depressive

AND ▾ All Fields ▾

Search or Add to history

History Download history Clear history

Search	Add to builder	Query	Items found	Time
#5	Add	Search **4[s_discriminator] AND (phs000007.v26.p10[s_ancestor] AND (4[s_discriminator] OR (1[s_discriminator] AND 1[s_has_analysis])))**	5624	13:54:12
#6	Add	Search **"macular degeneration" AND 1[Has Analysis] AND 1[s_discriminator]**	5	10:48:24

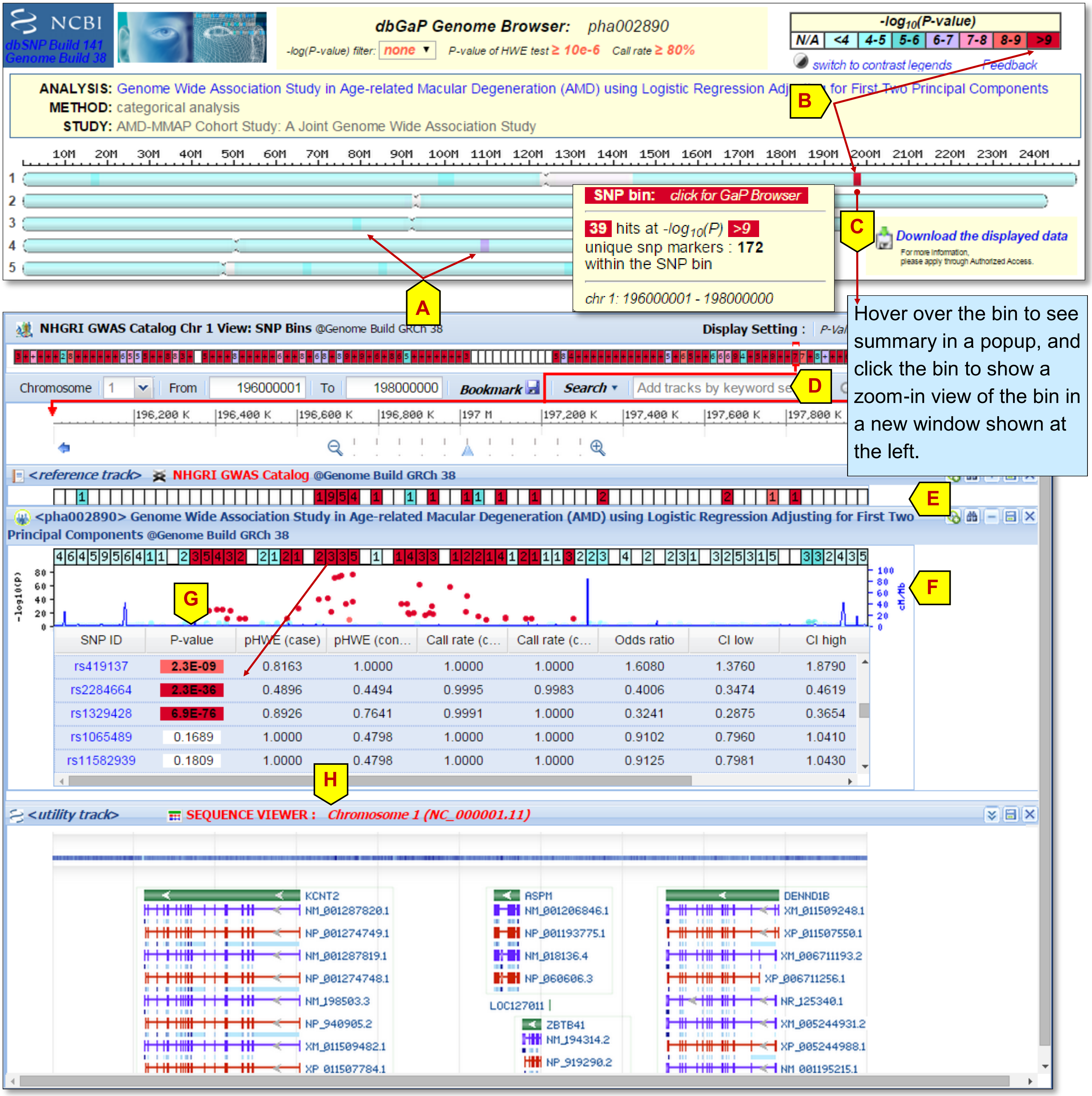
Studies (2) Variables (0) Documents (0) **Analyses (4)** Datasets (0)

Analysis	Description
Genome Wide Association Study in Age-related Macular Degeneration (AMD) using Logistic Regression Adjusting for First Two Principal Components Age related Macular Degeneration - MMAP Cohort: Association and Sequencing Studies Browse genome for pha002890.1	G -wide genotyping data were produced using the Illumina HumanCNV370v1_C array platform and filtered using the following criteria: 1) per sample call rate ≥ 99%, 2) SNP minor allele frequency ≥ 1%, 3) SNP call rate ≥ 95%, 4) SNP HWE p-value ≥ 1x10 -6 , 5) removing unexpected relatives, resulting in a set of 324067 autosomal SNPs across 3307 unrelated samples (2157 cases and 1150 controls).

Graphic Summary Through “Browse Genome for …”

“Browse genome for …” is linked to a graphical summary of the analysis results (shown below). Here, chromosomes are divided into bins of fixed sizes (**A**). These regions are color-coded according to the significance of the association between the identified genotype and the phenotype being studied, with bins in red indicating strong association (**B**). Clicking a bin opens a more detailed GaP Browser display for the region with a gene level resolution (**C**) or higher. In that display, the region shown is marked by the coordinates (**D**), which is further divided into smaller bins of fixed-length. A summary of sub-regions with significant phenotype-associated genotypes and number of SNPs found is displayed in the GWAS catalog track (**E**), while results from a specific analysis is shown in a track below (**F**), with SNPs for the region displayed in a scatter plot just below. Hovering-over a bin highlights the SNPs present in the region (**G**). The genome annotation pane (**H**), located under and aligned to the GWAS track, highlights the gene features found for the region. Clicking a bin in the track under the GWAS (**F**) zooms in to a much more detailed display for both panes (not shown). Displays can also be adjusted using controls at the top and in the left sidebar. Hovering over a control will display the online help.

https://www.ncbi.nlm.nih.gov/projects/SNP/gViewer/gView.cgi?aid=2890



NCBI dbSNP Build 141 Genome Build 38

dbGaP Genome Browser: *pha002890*

-log(P-value) filter: none ▼ P-value of HWE test ≥ 10e-6 Call rate ≥ 80%

-log10(P-value)							
N/A	<4	4-5	5-6	6-7	7-8	8-9	>9

switch to contrast legends Feedback

ANALYSIS: Genome Wide Association Study in Age-related Macular Degeneration (AMD) using Logistic Regression Adj... for First Two Principal Components
METHOD: categorical analysis
STUDY: AMD-MMAP Cohort Study: A Joint Genome Wide Association Study

SNP bin: *click for GaP Browser*
39 hits at -log10(P) **>9**
unique snp markers : **172**
within the SNP bin
chr 1: 196000001 - 198000000

Download the displayed data
For more information, please apply through Authorized Access.

Hover over the bin to see summary in a popup, and click the bin to show a zoom-in view of the bin in a new window shown at the left.

NHGRI GWAS Catalog Chr 1 View: SNP Bins @Genome Build GRCh 38 Display Setting : P-Val

Chromosome 1 ▼ From 196000001 To 198000000 Bookmark Search ▼ Add tracks by keyword se...

<reference track> **NHGRI GWAS Catalog** @Genome Build GRCh 38

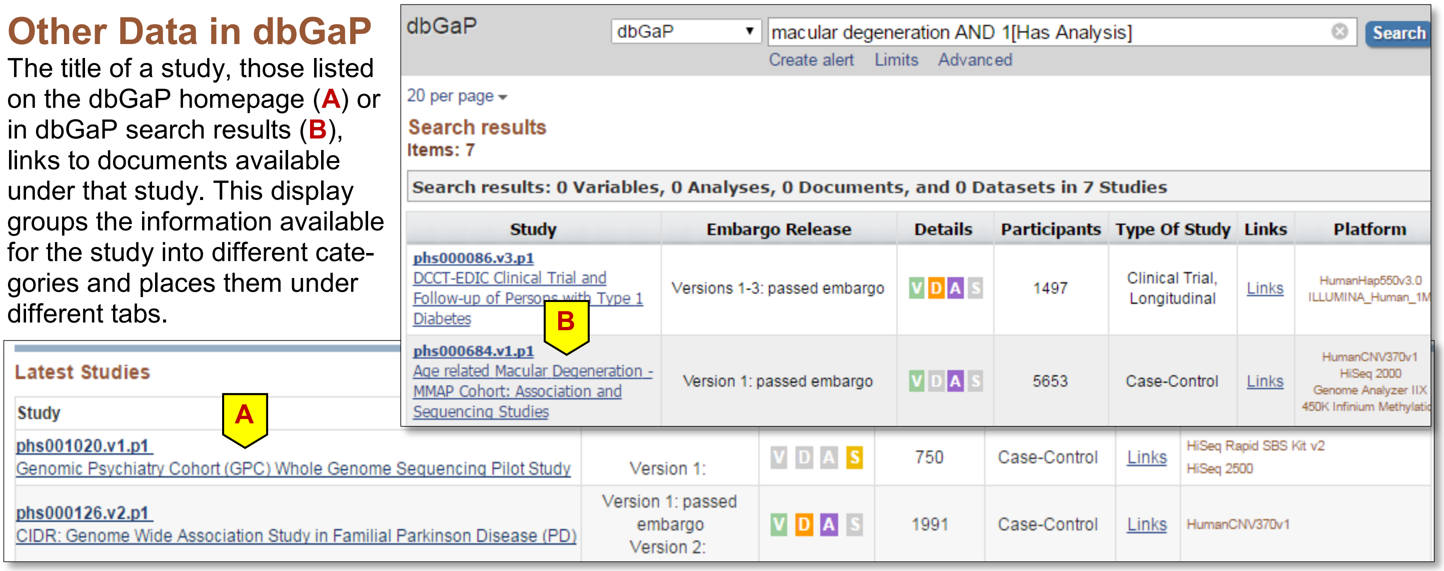
<pha002890> Genome Wide Association Study in Age-related Macular Degeneration (AMD) using Logistic Regression Adjusting for First Two Principal Components @Genome Build GRCh 38

SNP ID	P-value	pHWE (case)	pHWE (con...	Call rate (c...	Call rate (c...	Odds ratio	CI low	CI high
rs419137	2.3E-09	0.8163	1.0000	1.0000	1.0000	1.6080	1.3760	1.8790
rs2284664	2.3E-36	0.4896	0.4494	0.9995	0.9983	0.4006	0.3474	0.4619
rs1329428	6.9E-76	0.8926	0.7641	0.9991	1.0000	0.3241	0.2875	0.3654
rs1065489	0.1689	1.0000	0.4798	1.0000	1.0000	0.9102	0.7960	1.0410
rs11582939	0.1809	1.0000	0.4798	1.0000	1.0000	0.9125	0.7981	1.0430

<utility track> **SEQUENCE VIEWER :** *Chromosome 1 (NC_000001.11)*

Other Data in dbGaP

The title of a study, those listed on the dbGaP homepage (**A**) or in dbGaP search results (**B**), links to documents available under that study. This display groups the information available for the study into different categories and places them under different tabs.



dbGaP dbGaP ▾ macular degeneration AND 1[Has Analysis] Search

Create alert Limits Advanced

20 per page ▾

Search results

Items: 7

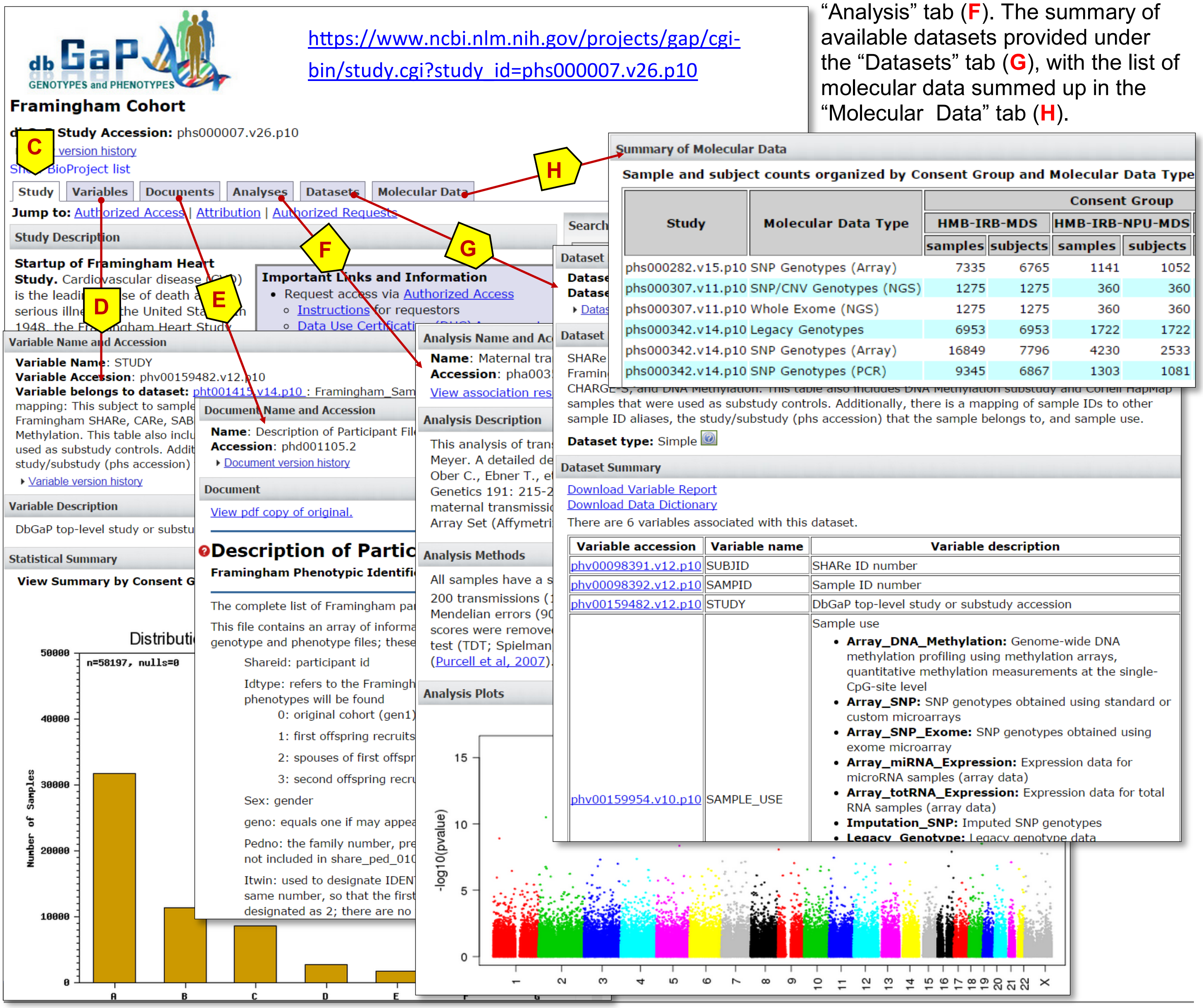
Search results: 0 Variables, 0 Analyses, 0 Documents, and 0 Datasets in 7 Studies

Study	Embargo Release	Details	Participants	Type Of Study	Links	Platform
phs000086.v3.p1 DCCT-EDIC Clinical Trial and Follow-up of Persons with Type 1 Diabetes	Versions 1-3: passed embargo	V D A S	1497	Clinical Trial, Longitudinal	Links	HumanHap550v3.0 ILLUMINA_Human_1M
phs000684.v1.p1 Age related Macular Degeneration - MMAP Cohort: Association and Sequencing Studies	Version 1: passed embargo	V D A S	5653	Case-Control	Links	HumanCNV370v1 HiSeq 2000 Genome Analyzer IIX 450K Infinium Methylatic

Latest Studies

Study						
phs001020.v1.p1 Genomic Psychiatry Cohort (GPC) Whole Genome Sequencing Pilot Study	Version 1:	V D A S	750	Case-Control	Links	HiSeq Rapid SBS Kit v2 HiSeq 2500
phs000126.v2.p1 CIDR: Genome Wide Association Study in Familial Parkinson Disease (PD)	Version 1: passed embargo Version 2:	V D A S	1991	Case-Control	Links	HumanCNV370v1

The “Study” tab (**C**) provides a general description on the goal of the study. The variables measured are listed under the “Variables” tab (**D**). Study documents with detailed background information and rationale for conducting the study are under the “Documents” tab (**E**). A summary of the analysis result with link to genomic display is given under the “Analysis” tab (**F**). The summary of available datasets provided under the “Datasets” tab (**G**), with the list of molecular data summed up in the “Molecular Data” tab (**H**).



https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v26.p10

Framingham Cohort

dbGaP Study Accession: phs000007.v26.p10

version history

BioProject list

Study | Variables | Documents | Analyses | Datasets | Molecular Data

Jump to: Authorized Access | Attribution | Authorized Requests

Study Description

Startup of Framingham Heart Study. Cardiovascular disease (CVD) is the leading cause of death and serious illness in the United States. In 1948, the Framingham Heart Study

Important Links and Information

- Request access via Authorized Access
  - Instructions for requestors
  - Data Use Certification

Variable Name and Accession

Variable Name: STUDY

Variable Accession: phv00159482.v12.p10

Variable belongs to dataset: pht001415.v14.p10 : Framingham_Sam

mapping: This subject to sample

Framingham SHARe, CARe, SAB

Methylation. This table also inclu

used as substudy controls. Addit

study/substudy (phs accession)

Variable version history

Variable Description

DbGaP top-level study or substu

Statistical Summary

View Summary by Consent G

Document Name and Accession

Name: Description of Participant Fil

Accession: phd001105.2

Document version history

Document

View pdf copy of original.

Description of Partic

Framingham Phenotypic Identifi

The complete list of Framingham pa

This file contains an array of informa genotype and phenotype files; these

Shareid: participant id

Idtype: refers to the Framingh phenotypes will be found

0: original cohort (gen1)

1: first offspring recruits

2: spouses of first offspr

3: second offspring recru

Sex: gender

geno: equals one if may appea

Pedno: the family number, pre not included in share_ped_010

Itwin: used to designate IDEN same number, so that the first designated as 2; there are no

Analysis Name and Ac

Name: Maternal tra

Accession: pha003

View association res

Analysis Description

This analysis of tran Meyer. A detailed de Ober C., Ebner T., e Genetics 191: 215-2 maternal transmissi Array Set (Affymetri

Analysis Methods

All samples have a s 200 transmissions (Mendelian errors (90 scores were remove test (TDT; Spielman (Purcell et al, 2007).

Analysis Plots

Search

Dataset

Datase

Datase

Datas

Dataset

SHARe

Framin

CHARGE-S, and DNA Methylation. This table also includes DNA Methylation substudy and Coriell HapMap samples that were used as substudy controls. Additionally, there is a mapping of sample IDs to other sample ID aliases, the study/substudy (phs accession) that the sample belongs to, and sample use.

Dataset type: Simple

Dataset Summary

Download Variable Report

Download Data Dictionary

There are 6 variables associated with this dataset.

Variable accession	Variable name	Variable description
phv00098391.v12.p10	SUBJID	SHARe ID number
phv00098392.v12.p10	SAMPID	Sample ID number
phv00159482.v12.p10	STUDY	DbGaP top-level study or substudy accession
phv00159954.v10.p10	SAMPLE_USE	Sample use • **Array_DNA_Methylation:** Genome-wide DNA methylation profiling using methylation arrays, quantitative methylation measurements at the single-CpG-site level • **Array_SNP:** SNP genotypes obtained using standard or custom microarrays • **Array_SNP_Exome:** SNP genotypes obtained using exome microarray • **Array_miRNA_Expression:** Expression data for microRNA samples (array data) • **Array_totRNA_Expression:** Expression data for total RNA samples (array data) • **Imputation_SNP:** Imputed SNP genotypes • **Legacy_Genotype:** Legacy genotype data

Summary of Molecular Data

Sample and subject counts organized by Consent Group and Molecular Data Type

Study	Molecular Data Type	Consent Group			
		HMB-IRB-MDS		HMB-IRB-NPU-MDS	
		samples	subjects	samples	subjects
phs000282.v15.p10	SNP Genotypes (Array)	7335	6765	1141	1052
phs000307.v11.p10	SNP/CNV Genotypes (NGS)	1275	1275	360	360
phs000307.v11.p10	Whole Exome (NGS)	1275	1275	360	360
phs000342.v14.p10	Legacy Genotypes	6953	6953	1722	1722
phs000342.v14.p10	SNP Genotypes (Array)	16849	7796	4230	2533
phs000342.v14.p10	SNP Genotypes (PCR)	9345	6867	1303	1081